Hindawi Publishing Corporation
PPAR Research
Volume 2016, Article ID 6042162, 6 pages
http://dx.doi.org/10.1155/2016/6042162

Research Article

PPARgene: A Database of Experimentally Verified and Computationally Predicted PPAR Target Genes

Li Fang,[1] Man Zhang,[1] Yanhui Li,[1] Yan Liu,[1] Qinghua Cui,[2] and Nanping Wang[1,3]

[1]*Institute of Cardiovascular Sciences, Peking University Health Science Center, Beijing 100191, China*
[2]*Department of Biomedical Informatics, Peking University Health Science Center, Beijing 100191, China*
[3]*The Advanced Institute for Medical Sciences, Dalian Medical University, Dalian 116044, China*

Correspondence should be addressed to Nanping Wang; nanpingwang2003@yahoo.com

Received 21 January 2016; Accepted 24 March 2016

Academic Editor: Todd Leff

The peroxisome proliferator-activated receptors (PPARs) are ligand-activated transcription factors of the nuclear receptor superfamily. Upon ligand binding, PPARs activate target gene transcription and regulate a variety of important physiological processes such as lipid metabolism, inflammation, and wound healing. Here, we describe the first database of PPAR target genes, PPARgene. Among the 225 experimentally verified PPAR target genes, 83 are for PPARα, 83 are for PPARβ/δ, and 104 are for PPARγ. Detailed information including tissue types, species, and reference PubMed IDs was also provided. In addition, we developed a machine learning method to predict novel PPAR target genes by integrating *in silico* PPAR-responsive element (PPRE) analysis with high throughput gene expression data. Fivefold cross validation showed that the performance of this prediction method was significantly improved compared to the *in silico* PPRE analysis method. The prediction tool is also implemented in the PPARgene database.

1. Introduction

Peroxisome proliferator-activated receptors (PPARs) are ligand-activated transcription factors that belong to the superfamily of nuclear receptors. PPARs form heterodimers with a retinoid X receptor (RXR) and control gene expression by binding to specific PPAR-responsive elements (PPREs) on target gene promoters [1]. PPARs play critical roles in the regulation of lipid and glucose metabolism, inflammation, wound healing, and many other pathophysiological processes [2–5]. Synthetic PPAR ligands, such as fibrates and thiazolidinediones, are used for clinical treatment of dyslipidemia and type 2 diabetes, respectively [6].

Extensive studies have demonstrated a variety of target genes regulated by the individual PPAR subtype. Therefore, building a database with a comprehensive collection of the previously verified PPAR target genes for each subtype will be helpful for PPAR research. In this study, we first established a database of PPAR target genes, PPARgene. Experimentally verified PPAR target genes were manually curated and detailed information including PPAR subtype, tissue types, species, and reference PubMed IDs was provided.

Recently, the application of high throughput technologies such as microarray has generated a number of PPAR-induced gene expression data sets, which are freely available in public database. By integrating *in silico* PPRE analysis with high throughput gene expression data, we developed a machine learning method to predict novel PPAR target genes. The prediction tool is also implemented in the PPARgene database (http://www.ppargene.org/).

2. Methods

2.1. Data Collection

2.1.1. Collection of Experimentally Verified PPAR Target Genes. PPAR-related publications were acquired from PubMed database using the key words "PPAR", "PPAR alpha", "PPAR beta", "PPAR delta", "PPAR gamma", or "peroxisome proliferator" (review articles were excluded). We then curated

the data manually and retrieved the PPAR target genes if experimental evidence for gene regulation (at mRNA and/or protein levels) and functional PPRE (reporter assay and/or DNA-binding assays) were both reported.

2.1.2. Collection of PPAR-Relevant Microarray Data Sets. PPAR-relevant microarray data sets were acquired by searching the GEO database [7] using the key words "PPAR", "PPAR alpha", "PPAR beta", "PPAR delta", "PPAR gamma", or "peroxisome proliferator". We manually curated 22 data sets in which PPARs were activated or overexpressed.

2.2. Feature Extraction

2.2.1. High Throughput Evidence (HTE). To obtain the high throughput experimental evidence supporting PPAR target gene interactions, we collected microarrays in which PPARs were activated or overexpressed. Raw data of collected microarrays were processed using the R-packages Bioconductor [8]. The HTE value of a gene was defined as total number of data sets divided by number of data sets in which this gene was upregulated ($\log_2$ fold change > 0.5).

2.2.2. PPRE Score (PS). Reference genome of mouse (GRCm38) and rat (Rnor_6.0) was downloaded from NCBI. According to previous studies [9–12], PPREs were located within 5 kb upstream or downstream of the transcription start site (TSS) in most cases. Therefore, we extracted −5 kb~+5 kb TSS flanking sequences from the reference genome for all mouse and rat genes identifiable by Entrez Gene ID according to the genomic coordinates.

Potential PPREs were scanned *in silico* using the position weight matrix (PWM) model, which was widely used to describe cis-regulatory elements [13, 14]. Since the three subtypes of PPARs bind to a common core consensus sequence, we did not distinguish the difference of binding site among subtypes and used the position frequency matrix (PFM) of PPARγ-RXRα heterodimer retrieved from JASPAR database (ID: MA0065.2) [15] to compute the PWM of PPRE. The PWM was computed as described previously [16]. Briefly, we calculated the PWM value as

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}, \tag{1}$$

where $W_{b,i}$ is PWM value of base b in position i, $p(b)$ is background probability of base b in the genome, and $p(b,i)$ is probability of base b in position i. Pseudocount values (square root of the number of sites) were added to each base in each position to smoothen the small sample effects. The PWM score for a putative sequence was calculated as sum of the PWM values for each nucleotide in the sequence. For each gene identifiable by Entrez Gene ID in the mouse genome, we scanned putative PPREs from the TSS flanking sequences in both strands at a PWM score cut-off of 4.56 (70% relative to top PWM score) initially. The PS value of a gene was defined as the highest PWM score of all PPREs identified in this gene.

2.2.3. Conserved PPRE Score (CPS). Evolutionary conservation has been used as an effective filter for improving specificity in regulatory motif recognition [17–19]. We performed comparative genomic analysis to identify conserved PPREs. Pairs of orthologous genes in mouse and rat were retrieved from NCBI HomoloGene database. TSS flanking sequences (−5 kb~+5 kb) of the orthologous gene pairs were aligned using megaBLAST with default parameters (word size = 28, reward = 1, mismatch penalty = −2, gap opening penalty = 0, and gap extension penalty = 2.5) [20, 21]. Alignments less than 50 bp or with an E-value > 0.001 were discarded. For each orthologous gene, we scanned putative PPREs from the TSS flanking sequences at a PWM score cut-off of 4.56. A pair of putative PPREs was identified as conserved PPRE if they were matched in the pairwise alignments. The CPS value of a gene was defined as the highest PWM score of all conserved PPREs identified in this gene.

2.3. Model Training and Evaluation

2.3.1. Training Sets for the Prediction Model. Experimentally verified target genes collected in the PPARgene database were defined as positive training samples. However, it would be difficult to prove that a gene is not a target gene of PPARs in any conditions. Thus, we obtain negative training samples by randomly choosing equal number of genes from the background data set, which contained all protein coding genes excluding the positive samples. To avoid sampling bias, we sampled the negative data set 100 times and then combined each negative data set with the positive data set to train the classifier.

2.3.2. Logistic Regression Classifier. We employed the binomial logistical regression model to predict PPAR target genes. All mouse protein coding genes with a HomoloGene database ID were classified according to a combination of the features described above. Let p_i be the probability that the ith gene is a PPAR target gene and let $1 - p_i$ be the probability that it is not. The logistic regression model is

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \sum_{j=1}^{M} \beta_j X_{ij}, \tag{2}$$

where β_j is the regression coefficient of the feature X_{ij}. The logistic regression model was implemented using the generalized linear model (GLM) function in R [22].

2.3.3. Performance Evaluation. We used 5-fold cross validation to evaluate the performance of the logistic regression model. In each round, 20% of the samples were left out as the test data and the remaining were the training data. Precision, recall, and $F1$ score were used to evaluate the performance

of the classifier. Precision, recall, and $F1$ were calculated as

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})},$$
$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \quad (3)$$
$$F1 = \frac{2 \text{ precision} \times \text{recall}}{(\text{precision} + \text{recall})},$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. We also calculated AUC, the area under the receiver operating characteristic (ROC) curve, using ROCR package [23]. Because negative data sets were obtained by 100 random samplings, the medians of precisions, recalls, $F1$s, and AUCs of the 100 training results were used.

2.4. Web Server. All data were organized using MySQL, an open-source relational database management system. The website was presented using PHP. The PPARgene database is freely available at http://www.ppargene.org/.

3. Results and Discussion

3.1. Experimentally Verified PPAR Target Genes. In this study, we developed a database for PPAR target genes. We curated PPAR target genes manually from 9046 PPAR-related publications. The PPARgene database now contains 225 experimentally verified PPAR target genes, including 83 target genes for PPARα, 83 target genes for PPARβ/δ, and 104 genes for PPARγ. Forty genes were common targets of at least two PPAR subtypes. Detailed information including tissues, species, reference PubMed IDs, and hyperlinks to the original articles in PubMed database was also provided.

3.2. Generation of Logistic Regression Models to Predict PPAR Target Genes. We generated a logistic regression model to predict novel PPAR target genes. To train the logistic regression model, experimentally verified target genes were used as positive examples. Equal numbers of negative examples were obtained by random sampling from the background gene sets. Since the three PPAR subtypes bind to a conserved core sequence and share some common target genes [24], we currently did not distinguish subtypes in our prediction model.

Firstly, we generated the prediction model only based on *in silico* PPRE recognition using the standard position weight matrices (PWM) model [16]. Because functional PPREs were also found in downstream region of the TSS [9–12, 25], we scanned PPREs on both upstream and downstream regions. Genes were predicted as target genes or not according to the PWM score (PS). Fivefold cross validation was used to evaluate the performance of this model. As shown in Table 1, the median precision, recall, $F1$, and AUC were 0.57, 0.49, 0.52, and 0.59, respectively. The performance was poor, which may be due to a high number of false predictions of PPREs.

Table 1: Performances of logistic regression models trained on different features.

Features	Precision	Recall	$F1$	AUC
PS	0.57	0.49	0.52	0.59
CPS	0.61	0.68	0.64	0.68
CPS + HTE	0.83	0.59	0.69	0.82

PS: PPRE score; CPS: conserved PPRE score; HTE: high throughput evidence.

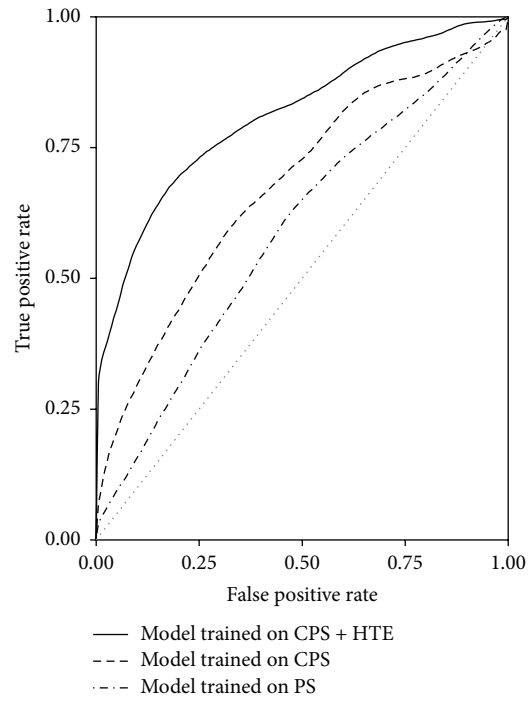


Figure 1: ROC curves for logistic regression models trained on different features. CPS: conserved PPRE score; HTE: high throughput evidence; PS: PPRE score.

It is reported that conservation in regulatory regions can be used to enhance the predictive specificity [17–19]. We next performed comparative genomic analysis to identify putative PPREs conserved in mouse and rat. Orthologous genes were then classified according to conserved PPRE score (CPS). As shown in Table 1, the median precision, recall, $F1$, and AUC were 0.61, 0.68, 0.64, and 0.68, respectively, which indicated a better performance.

Rather than *in silico* prediction of binding sites, experimental data sets provide direct evidence for gene regulation. Recently, high throughput technologies have produced a number of public available PPAR-relevant gene expression profiles. Thus, we collected PPAR-gain-of-function microarray data sets from the GEO database and extracted the supporting evidence for gene regulation. The logistic regression model was then generated based on a combination of conserved PPRE score and high throughput evidence. As shown in Table 1, the median precision, recall, $F1$, and AUC were 0.61, 0.68, 0.64, and 0.68. The performance was greatly improved. ROC curves of the prediction models also showed the improvement in performance (Figure 1).

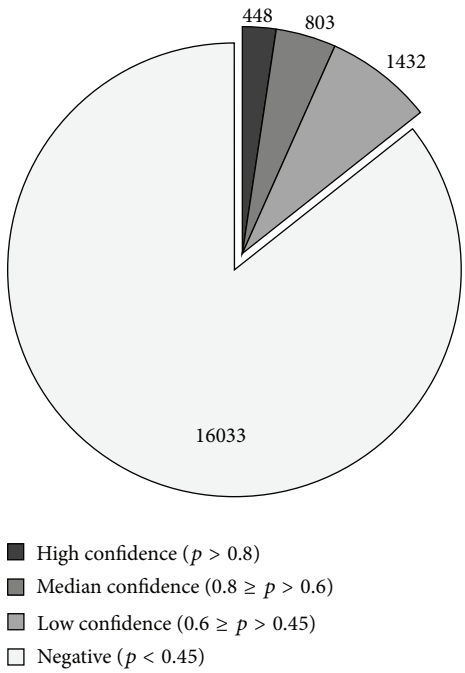


Figure 2: Number of predicted target genes in mouse genome. The predicted target genes were classified into 3 confidence levels according to the p value computed in the logistic regression model.

3.3. Genome-Wide Prediction of PPAR Target Genes. We predicted PPAR target genes from all 18,716 orthologous genes in mouse genome using the prediction model based on the combination of conserved PPRE score and high throughput evidence. We classified the predicted target genes into 3 confidence levels according to the p value (the probability of being a PPAR target gene) (Figure 2). In total, 2,683 genes with $p > 0.45$ were predicted as potential PPAR target genes, in which 448 genes were in the high-confidence category ($p > 0.8$), 803 genes were in the median-confidence category ($0.8 \geq p > 0.6$), and 1432 genes were in the low-confidence (high-sensitivity) category ($0.6 \geq p > 0.45$). Genes with p value ≤ 0.45 were predicted as negative. A complete list of the predicted PPAR target genes was available in the PPARgene website.

4. Querying the Database

The PPARgene database is composed of two modules: one is for querying experimentally verified target genes and the other is for querying computationally predicted target genes.

4.1. Experimentally Verified Target Genes. We provide users two ways to query the experimentally verified target genes. First, users can browse the results by selecting the PPAR subtype. PPARgene will return a table of matched entries. Users can also submit a specific gene symbol. The provided results contain the following items: PPAR subtype, gene symbol, species, tissue/cell types, regulation direction, and reference PubMed IDs.

4.2. Computationally Predicted Target Genes. Users can retrieve the prediction results by querying the gene symbol. If the gene is predicted as a PPAR target gene, the query will return a p value with a confidence level. A larger p value means a higher confidence. High throughput gene expression data and putative PPREs were listed to support the prediction. For example, Klf15 was predicted as a PPAR target gene at a high confidence (Figure 3). The prediction was made based on the curated microarray data and identified PPREs. PPAR agonists WY14643 and GW501516 upregulated Klf15 expression in mouse heart and skeletal muscle tissues. In addition, 9 putative PPREs were found in the TSS flanking regions of mouse Klf15. Six of the 9 PPREs were also found in rat Klf15 and labeled with an asterisk. The PPRE in the +1102 has a highest PWM score (13.45). Thus, the logistic regression model integrated both the gene expression information and the highest PWM score of the PPRE to compute the probability value (p) as 0.84298, which placed Klf15 as a predicted target gene in the high-confidence category.

4.3. Downloadable Files. Users can download data sets of experimentally verified PPAR target genes as well as computationally predicted target genes. We also provide hyperlinks for downloading the high throughput experimental data sets curated in our prediction model.

5. Future Extensions

In this release of PPARgene, we have focused on curation and prediction of protein coding target genes. Recent studies demonstrated that PPARs regulate non-protein coding genes as well [26, 27]. Therefore, the future goal is to predict noncoding target genes of PPARs. We will also develop methods to predict target genes for each PPAR subtype. Experimentally supported PPAR target genes in the PPARgene database will be updated every 3 months.

6. Conclusion

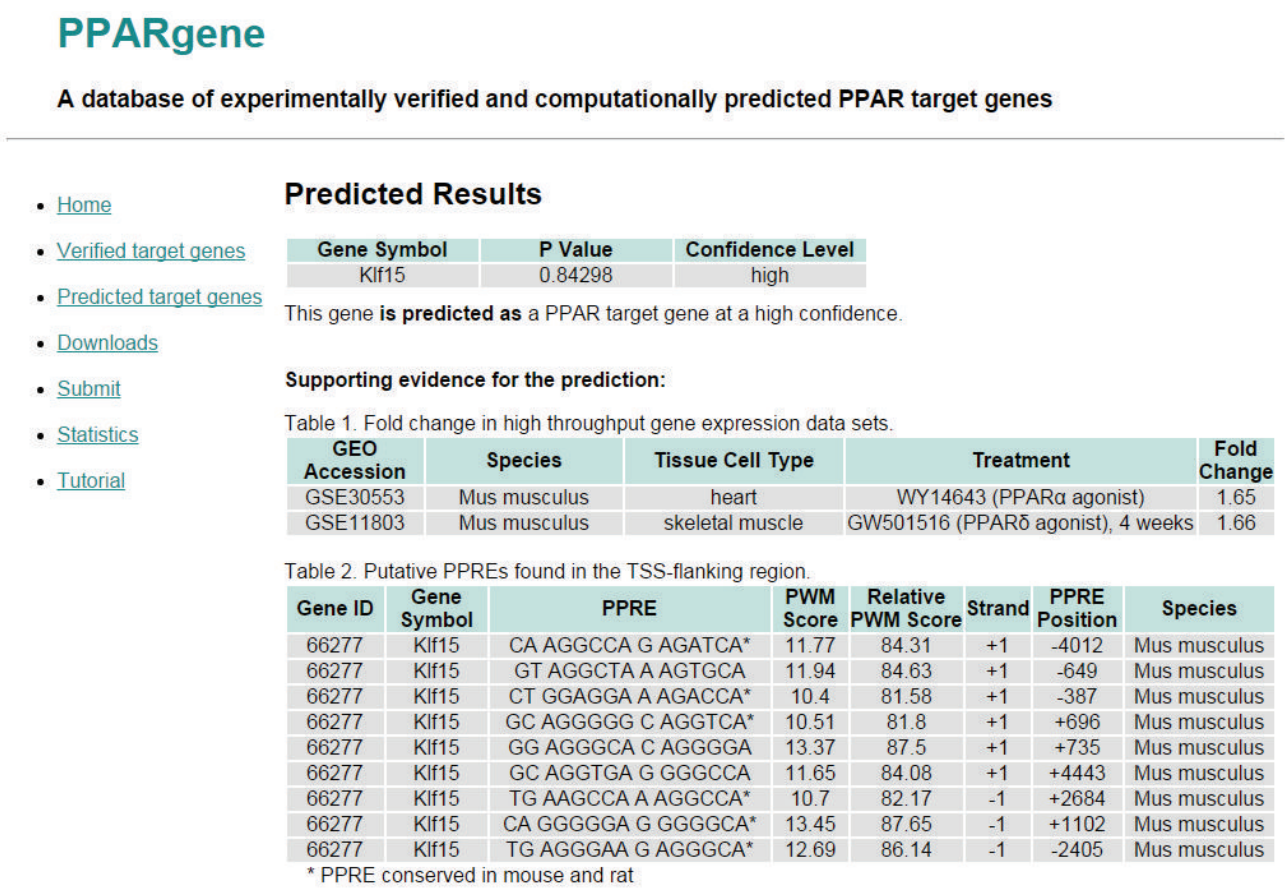
In this study, we described PPARgene, a novel database of experimentally verified as well as computationally predicted PPAR target genes. By integrating *in silico* PPRE analysis with high throughput gene expression data, we developed an effective machine learning method to predict novel PPAR target genes in the mouse genome. We consider that PPARgene will be a useful tool for PPAR research.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was funded by grants from the National Science Foundation of China (31430045, 81470373, and 81220108005).



PPARgene

A database of experimentally verified and computationally predicted PPAR target genes

- Home
- Verified target genes
- Predicted target genes
- Downloads
- Submit
- Statistics
- Tutorial

Predicted Results

Gene Symbol	P Value	Confidence Level
Klf15	0.84298	high

This gene **is predicted as** a PPAR target gene at a high confidence.

Supporting evidence for the prediction:

Table 1. Fold change in high throughput gene expression data sets.

GEO Accession	Species	Tissue Cell Type	Treatment	Fold Change
GSE30553	Mus musculus	heart	WY14643 (PPARα agonist)	1.65
GSE11803	Mus musculus	skeletal muscle	GW501516 (PPARδ agonist), 4 weeks	1.66

Table 2. Putative PPREs found in the TSS-flanking region.

Gene ID	Gene Symbol	PPRE	PWM Score	Relative PWM Score	Strand	PPRE Position	Species
66277	Klf15	CA AGGCCA G AGATCA*	11.77	84.31	+1	-4012	Mus musculus
66277	Klf15	GT AGGCTA A AGTGCA	11.94	84.63	+1	-649	Mus musculus
66277	Klf15	CT GGAGGA A AGACCA*	10.4	81.58	+1	-387	Mus musculus
66277	Klf15	GC AGGGGG C AGGTCA*	10.51	81.8	+1	+696	Mus musculus
66277	Klf15	GG AGGGCA C AGGGGA	13.37	87.5	+1	+735	Mus musculus
66277	Klf15	GC AGGTGA G GGGCCA	11.65	84.08	+1	+4443	Mus musculus
66277	Klf15	TG AAGCCA A AGGCCA*	10.7	82.17	-1	+2684	Mus musculus
66277	Klf15	CA GGGGGA G GGGGCA*	13.45	87.65	-1	+1102	Mus musculus
66277	Klf15	TG AGGGAA G AGGGCA*	12.69	86.14	-1	-2405	Mus musculus

* PPRE conserved in mouse and rat

Figure 3: Predicted results of a query gene. High throughput gene expression data and putative PPREs were provided to support the prediction.

References

[1] J. Berger and D. E. Moller, "The mechanisms of action of PPARs," *Annual Review of Medicine*, vol. 53, pp. 409–435, 2002.

[2] Y. Fan, Y. Wang, Z. Tang et al., "Suppression of pro-inflammatory adhesion molecules by PPAR-δ in human vascular endothelial cells," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 28, no. 2, pp. 315–321, 2008.

[3] F. S. Harman, C. J. Nicol, H. E. Marin, J. M. Ward, F. J. Gonzalez, and J. M. Peters, "Peroxisome proliferator-activated receptor-delta attenuates colon carcinogenesis," *Nature Medicine*, vol. 10, no. 5, pp. 481–483, 2004.

[4] N. Wang, L. Verna, N.-G. Chen et al., "Constitutive activation of peroxisome proliferator-activated receptor-γ suppresses pro-inflammatory adhesion molecules in human vascular endothelial cells," *Journal of Biological Chemistry*, vol. 277, no. 37, pp. 34176–34181, 2002.

[5] Y.-X. Wang, "PPARs: diverse regulators in energy metabolism and metabolic diseases," *Cell Research*, vol. 20, no. 2, pp. 124–137, 2010.

[6] B. Staels and J.-C. Fruchart, "Therapeutic roles of peroxisome proliferator-activated receptor agonists," *Diabetes*, vol. 54, no. 8, pp. 2460–2470, 2005.

[7] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D991–D995, 2013.

[8] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, article R80, 2004.

[9] M. I. Lefterova, Y. Zhang, D. J. Steger et al., "PPARγ and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale," *Genes and Development*, vol. 22, no. 21, pp. 2941–2952, 2008.

[10] R. Nielsen, T. Å. Pedersen, D. Hagenbeek et al., "Genome-wide profiling of PPARγ:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis," *Genes and Development*, vol. 22, no. 21, pp. 2953–2967, 2008.

[11] D. L. M. van der Meer, T. Degenhardt, S. Väisänen et al., "Profiling of promoter occupancy by PPARα in human hepatoma cells via ChIP-chip analysis," *Nucleic Acids Research*, vol. 38, no. 9, pp. 2839–2850, 2010.

[12] T. Adhikary, A. Wortmann, T. Schumann et al., "The transcriptional PPARβ/δ network in human macrophages defines a unique agonist-induced activation state," *Nucleic Acids Research*, vol. 43, no. 10, pp. 5033–5051, 2015.

[13] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.

[14] M. L. Bulyk, "Computational prediction of transcription-factor binding site locations," *Genome Biology*, vol. 5, no. 1, article 201, 2003.

[15] A. Mathelier, X. Zhao, A. W. Zhang et al., "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. 42, no. 1, pp. D142–D147, 2014.

[16] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 276–287, 2004.

[17] A. Sandelin, W. W. Wasserman, and B. Lenhard, "ConSite: web-based prediction of regulatory elements using cross-species comparison," *Nucleic Acids Research*, vol. 32, pp. W249–W252, 2004.

[18] X. Xie, J. Lu, E. J. Kulbokas et al., "Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals," *Nature*, vol. 434, no. 7031, pp. 338–345, 2005.

[19] A. Stark, M. F. Lin, P. Kheradpour et al., "Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures," *Nature*, vol. 450, no. 7167, pp. 219–232, 2007.

[20] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 203–214, 2000.

[21] C. Camacho, G. Coulouris, V. Avagyan et al., "BLAST+: architecture and applications," *BMC Bioinformatics*, vol. 10, article 421, 2009.

[22] R. C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015.

[23] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.

[24] M. Rakhshandehroo, B. Knoch, M. Müller, and S. Kersten, "Peroxisome proliferator-activated receptor alpha target genes," *PPAR Research*, vol. 2010, Article ID 612089, 20 pages, 2010.

[25] A. Bugge, M. Siersbæk, M. S. Madsen, A. Göndör, C. Rougier, and S. Mandrup, "A novel intronic peroxisome proliferator-activated receptor γ enhancer in the Uncoupling Protein (UCP) 3 gene as a regulator of both UCP2 and -3 expression in adipocytes," *The Journal of Biological Chemistry*, vol. 285, no. 23, pp. 17310–17317, 2010.

[26] K.-J. Yin, Z. Deng, M. Hamblin et al., "Peroxisome proliferator-activated receptor δ regulation of miR-15a in ischemia-induced cerebral vascular endothelial injury," *Journal of Neuroscience*, vol. 30, no. 18, pp. 6398–6408, 2010.

[27] X. Fang, L. Fang, A. Liu, X. Wang, B. Zhao, and N. Wang, "Activation of PPAR-delta induces microRNA-100 and decreases the uptake of very low-density lipoprotein in endothelial cells," *British Journal of Pharmacology*, vol. 172, no. 15, pp. 3728–3736, 2015.